[redacted]

working paper
department
of economics

Credible Neologisms in Games of Communication

Joseph Farrell

No. 386

June 1985

massachusetts
institute of
technology
50 memorial drive
cambridge, mass. 02139

Credible Neologisms in Games of Communication

Joseph Farrell

No. 386 June 1985

Credible Neologisms in Games of Communication

by

Joseph Farrell

GTE Laboratories
40 Sylvan Road
Waltham, MA 02254

and M.I.T.

December 1983, revised June 1985

Thanks are due to many people for the much-delayed revision of this paper. Vincent Crawford, Jerry Green, Eric Maskin, Andrew McLennan, and especially Roger Myerson and Joel Sobel have contributed ideas and encouragement. A special debt is due to Suzanne Scotchmer and Joel Sobel, both of whom kept encouraging me. I also wish to thank Kathleen Stevens for programming assistance.

1. Introduction

Little attention has been paid to communication in games: for the most part, game theorists think as if no player would ever trust another (as is the case with two-person zero-sum games). But an important role of language is coordination of actions, and game theory is the natural discipline in which to study language from that point of view, as Lewis [9] showed.

In [3], we defined communication equilibrium in games, and showed that it is a strict and useful extension of Bayesian equilibrium. However, that notion, like that used by Crawford and Sobel [2] and Green and Stokey [4], requires only that given the way language is interpreted, there is no incentive to lie. In the present paper we extend the analysis by examining incentives to introduce credible neologisms.

This can be viewed in two ways. Our principal interpretation is that it is an equilibrium condition that nobody should have an incentive to introduce a credible neologism: for, if someone has, then the purported equilibrium does not describe how the game will be played. We elaborate on this below, when we define a neologism-proof equilibrium. Another related approach, which we only hint at here in Section 5, is to ask when an evolving language will develop neologisms that will not immediately be discredited by lies: a self-signaling neologism will be used truthfully, at least at first.

More technically, the paper can be viewed as examining the implications of a certain restriction on disequilibrium beliefs in sequential equilibrium. The restriction is that, if there is a set X of types that strictly prefer being treated as a group X to what they get in equilibrium, and if no other types would wish to treated as X, then it must be possible for the message-sender to induce the belief X.

This restriction is similar to that independently proposed by Grossman and Perry [5]. However, they, like the literature on stability of signaling equilibrium (Banks and Sobel [1], Kohlberg and Mertens [6], Kreps [7]) consider signals that directly affect payoffs. Their criteria, being based on dominance arguments, have no force if, as here, signals do not directly affect payoffs.

The paper is organized as follows. In Section 2, we define a simple communication game, and a sequential equilibrium in such a game. We introduce the notion of a self-signaling set of types (an X with the properties just described), and argue that if there is such a set then the equilibrium is unpersuasive, as it is vulnerable to the introduction of a credible neologism. An equilibrium not subject to such an objection we call _neologism-proof_.

In Section 3, we give four examples to illustrate our concept, and relate it to Myerson's notion [10] of core mechanism. We show by example that neologism-proof equilibrium need not exist.

In Section 4, we prove some existence results for neologism-proof equilibrium. In Section 5, we consider what happens if agents repeatedly play (adjusting towards best responses) a game that has no neologism-proof equilibrium. In our example, the average result depends on relative speeds of adjustment, and we suggest that the usual game-theoretic paradigm (in which all those speeds are assumed infinite) may be misleading. Section 6 concludes.

2. Assumptions and Definitions

2.1 Simple Communication Games

First, we define our object of analysis: a simple communication game G. There are two[1] players: the Sender (S) and the Receiver (R). S has payoff-relevant private information $t \in T$, which we sometimes call his type. We assume that T is finite, and that R's prior π on T, which is common knowledge, assigns strictly positive probability to each type t. The only choice that enters directly into payoffs is R's choice of an action $a \in A$, where A is a fixed finite set; payoffs also depend on t. Players maximize their expected payoffs $u^S(a,t)$ and $u^R(a,t)$.

The Sender, S, chooses a message m from a class M^* of possible messages; R then learns m before choosing a. We think of M^* as infinite but discrete - for instance, the set of all (arbitrarily long) utterances in English.

2.2 Sequential Equilibrium

A sequential equilibrium (Kreps and Wilson, [8]) of the game $G = (T, A, M^*, u^S(\bullet,\bullet), u^R(\bullet,\bullet), \pi)$ consists of the following:

(i) A function s: $T \rightarrow \Delta(M^*)$ which assigns to each $t \in T$ a probability distribution s(t) on M^*.

(ii) A function p: $M^* \rightarrow \Delta(T)$ which assigns to each $m \in M^*$ a probability distribution p(m) on T.

(iii) A function r: $M^* \rightarrow \Delta(A)$ which assigns to each $m \in M^*$ a probability distribution r(m) on A.

The functions s and r are S's and R's Bayesian strategies respectively. The distribution p(m) on T represents R's posterior beliefs about $t \in T$, after hearing the message m.

These functions must satisfy three conditions:

First, for each $t \in T$ and each $m \in M^*$,

$$\pi(t)s(t)(m) = p(m)(t)\sum_{t \in T}\pi(t)s(t)(m) \qquad (2.1)$$

This condition says that the probability distributions on $T \times M^*$ described by $\pi(\bullet)$ and $s(\bullet)$, and by $p(\bullet)$ and $s(\bullet)$, must be the same. In other words, given $s(\bullet)$, R's posteriors must be consistent with his priors on t. We can re-write (2.1) in more familiar Bayes' rule form by dividing by the prior probability of m, $q(m) \equiv \sum_{t\in T}\pi(t)s(t)(m)$, if $q(m) > 0$, but so as to allow for zero-probability messages we use the form (2.1), which can be re-written as:

$$\pi(t)s(t)(m) = q(m)p(m)(t) \qquad (2.2)$$

The second condition is that R's choice of action be optimal given his beliefs. Write $v^R(m) \equiv \sup_{a\in A}[\sum_{t\in T}p(m)(t)u^R(a,t)]$ for R's optimized expected payoff when he hears message m. Then we require that:

$$r(m)(a) > 0 \rightarrow \sum_{t\in T}p(m)(t)u^R(a,t) = v^R(m) \qquad (2.3)$$

Our third requirement is that, given R's response function $r(\bullet)$, S's choices of messages should be optimal. Define $v^S(t) \equiv \sup_{m\in M*}[\sum_{a\in A}r(m)(a)u^S(a,t)]$ to be S's optimized expected payoff when he observes t. Then we require[2]

$$s(t)(m) > 0 \rightarrow \sum_{a\in A}r(m)(a)u^S(a,t) = v^S(t) \qquad (2.4)$$

Having defined a sequential equilibrium, we prove[3] that a sequential equilibrium always exists:

Proposition 1: Every simple communication game has a sequential equilibrium.

Proof: Let $\bar{a} \in A$ be optimal for R given his prior beliefs π. Let $f(\bullet)$ be any probability distribution on M^*. Define:

$$s(t) = f(\bullet) \quad \text{for all } t \in T$$

$$p(m) = \pi(\bullet) \quad \text{for all } m \in M^*$$

$$r(m)(a) = \begin{cases} 1 & \text{if } a = \bar{a} \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } m \in M^*.$$

This is a sequential equilibrium. In this equilibrium, of course, no information about t is communicated: we call it the uncommunicative equilibrium.

2.3 Outcomes and Equivalence of Equilibria

Next, we define a notion of equivalence of equilibria. We are not really concerned with the choice of messages; rather, we care about how communication affects the outcome of G. We define the outcome w of an equilibrium (s,p,r) of G to be the function w from T to $\Delta(A)$ given by:

w(t)(a) = probability that a is chosen if t is the state;

$$= \sum_{m \in M^*} s(t)(m)r(m)(a) \qquad (2.5)$$

Two equilibria are <u>equivalent</u> if they have the same outcome. While there are always infinitely many sequential equilibria[4], there may be many fewer equivalence classes, and we often abuse language slightly by saying (e.g.) that (s,p,r) is unique, when we mean that every equilibrium is equivalent to (s,p,r).

2.4 Further Notation

Let X be a non-empty subset of T. We write $\pi(\bullet|X)$ for the probability distribution on T given by

$$\pi(t|X) = \begin{cases} \dfrac{\pi(t)}{\sum_{t' \in X} \pi(t')} & \text{if } t \in X \\ 0 & \text{otherwise} \end{cases}$$

We write $a^*(X)$ for R's optimal action if his beliefs are $\pi(\bullet|X)$, and <u>assume</u> that a*(X) is unique for each X. Since T and A are finite, this is generically true. It also holds if R's preferences are strictly concave in a, as in Crawford and Sobel [2].

We write v(X,t) for t's payoff if R has beliefs $\pi(\bullet|X)$:

$$v(X,t) \equiv u^S(a^*(X),t).$$

Given a response rule r on R's part, we define t's payoff:

$$v(r,t) \equiv \sup_{m \in M^*} \sum_{a \in A} r(m)(a) u^S(a,t)$$

and recall that, if r is part of a sequential equilibrium, then v(r,t) is achieved. We define the set P(X,r) of types who strictly prefer to be thought of as unknown members of X than to be treated according to r:

$$P(X,r) \equiv \{t \in T \mid v(X,t) > v(r,t)\} \qquad (2.6)$$

When there is no danger of confusion, we write P(X) for P(X,r).

2.5 Self-Signaling Subsets and Neologism-Proof Equilibrium

We say that the subset X is self-signaling if

$$P(X) = X \qquad (2.7)$$

Now suppose that in a sequential equilibrium (s,p,r), R confidently believes that S expects to be treated according to r; and suppose that, instead of sending a message sent with positive probability in equilibrium, S sends an unexpected message or neologism "t is in X." If (2.7) holds, then it seems reasonable to expect that R should believe that t is in X. For if $t \notin X$, then (by (2.7)) S would not try to persuade R that $t \epsilon X$. Moreover, every type t with $t \epsilon X$ would try to persuade R that $t \epsilon X$. Thus, we submit, a self-signaling message is highly compelling.

Based on this, we give the following definition: A sequential equilibrium (s,p,r) of a simple communication game is <u>neologism-proof</u> if there is no non-empty subset X of T that is self-signaling relative to the payoffs v(r,t).

A sequential equilibrium is a language (set of messages used in equilibrium, and their interpretations p) that will not induce lying. Our concept of neologism-proof equilibrium not only requires that (we start with a sequential equilibrium), but also requires that no credible neologisms (new messages, literally "new words") will be used.

There is always more language available than is used in a given game: non-equilibrium messages exist[5]. Moreover, these non-equilibrium messages have focal meanings: their literal meanings. Thus it is unreasonable to assume, as the concept of sequential equilibrium does, that R will not <u>understand</u> neologisms. We will assume that, for every non-empty subset X of T, there is a neologism n(X) that (it is common knowledge) claims that <u>t</u> is in X. We will focus on whether n(X) is <u>credible</u>.

While we certainly should not assume that R is credulous enough to believe arbitrary statements, we also should not assume that he will not believe a convincingly credible neologism. Because a self-signaling neologism is a message whose common-knowledge focal meaning is such that S would like R to believe it if and only if it is true, we assert that only a paranoid Receiver would refuse to believe a self-signaling neologism. Accordingly, we assume that, if a self-signaling neologism n(X) is available, then S can (and, if t is in X, will) make R believe that t is in X. Moreover, since all and only types $t \in X$ would use the neologism n(X), R's beliefs ought to be simply $\pi|X$. Therefore, S knows that by using n(X) he can induce beliefs $\pi|X$ and action a*(X). So it is a compelling restriction on sequential equilibrium that there be no self-signaling subset X.

Myerson [10] has discussed a related equilibrium concept (the "core mechanism") in which fewer neologisms are credible than in the present paper. He requires that a neologism (in his terms, a proposed alternative mechanism) be preferred by some set X of types, and be incentive-compatible (have R making a best response to his beliefs) whatever R's beliefs "between" π and $\pi|X$. This stronger requirement on neologisms makes fewer neologisms credible, and thus makes more sequential equilibria neologism-proof. Myerson obtains an existence theorem (his Theorem 4), while we show below in a very simple example that neologism-proof equilibrium need not exist.

3. Some Simple Examples

In this section, we analyse the sequential and the neologism-proof equilibria of some simple communication games. We show how the requirement of neologism-proofness eliminates many "unsatisfactory" sequential equilibria. We will also see that, even in non-pathological games, neologism-proof equilibrium may fail to exist.

Example 1: Eliminating an Uncommunicative Equilibrium

As we saw above, there is always a sequential equilibrium in which different types are treated alike, and any attempt to communicate is ignored, even if it is entirely credible. Example 1 shows how our condition of neologism-proofness may rule out this equilibrium.

We define a simple communication game as follows:

$T = \{t_1, t_2\}$

$\pi(t_1) = \pi(t_2) = 1/2$

$A = \{a_1, a_2, a_3\}$

Payoffs can be concisely described as follows:

S's payoff:

	t_1	t_2
a_1	1	-2
a_2	-2	-1
a_3	0	0

R's payoff:

	t_1	t_2
a_1	3	0
a_2	0	3
a_3	2	2

R's best responses in terms of his posterior probability p that S is t_1 are:

$$a^*(p) = \begin{cases} a_2 & \text{if } p \leq 1/3 \\ a_1 & \text{if } p \geq 2/3 \\ a_3 & \text{if } 1/3 \leq p \leq 2/3 \end{cases}$$

There are two sequential equilibria: the uncommunicative one, and one in which S reveals his type to R. The uncommunicative equilibrium is better (ex-ante) for S. However, even if S can propose behavior ex-ante, he cannot get that allocation, because if he turns out to be of type 1 he will use a self-signaling neologism. That is, $X = \{t_1\}$ is self-signaling relative to the payoff S gets in the uncommunicative equilibrium. Therefore that equilibrium is not neologism-proof. The revealing equilibrium is neologism-proof.

Example 2: Eliminating an Informative Equilibrium

This game also has an informative and an uninformative sequential equilibrium. However, in this case it is the informative equilibrium that fails to be neologism-proof.

The game is the same as in Example 1, except that $u^S(a_1,t_1)$ is changed from 1 to -1. Now S's payoffs are:

	t_1	t_2
a_1	-1	-2
a_2	-2	-1
a_3	0	0

If R responds to some equilibrium messages with a_1 and to all others with a_2, then truth-telling is a sequential equilibrium. However, it is vulnerable to the neologism n(T): "I won't tell you the state; since this is better (0 > -1) for t_1 and for t_2, you shouldn't infer anything about t." That is, X = T is self-signaling relative to payoffs resulting from revelation, and therefore the revealing equilibrium is not neologism-proof.

Example 3: No neologism-proof equilibrium

In this example, we produce a game that has no neologism-proof equilibrium.

As in examples 1 and 2, $T = \{t_1, t_2\}$, $\pi(t_1) = 1/2 = \pi(t_2)$, and $A = \{a_1, a_2, a_3\}$. R's payoffs are also as in examples 1 and 2. S's payoffs are given by:

	t_1	t_2
a_1	2	-1
a_2	-1	-2
a_3	0	0

There is now just one sequential equilibrium outcome: action a_3 is chosen with probability one in both states. But $\{t_1\}$ is self-signaling relative to equilibrium payoffs. Therefore, there is no neologism-proof equilibrium.

Example 4: The Crawford-Sobel Game

We now show that the example discussed by Crawford and Sobel ([2], Section 4) often fails to have a neologism-proof equilibrium, and never has a neologism-proof equilibrium with communication. The game has infinite type and action spaces, but concavity ensures that it is well-behaved. $T = [0,1]$, and the prior $\pi(\bullet)$ is uniform. The action space A is also $[0,1]$. Payoffs are

$$u^R(a,t) = -(a-t)^2$$
$$u^S(a,t) = -(a-(t+b))^2 \quad (b > 0)$$

Crawford and Sobel show that, if N(b) is the largest integer N satisfying $2N(N-1)b < 1$, then there is one sequential equilibrium for each integer $1 \leq N \leq N(b)$. The N-equilibrium is described by the partition

$$[0,1] = [X_0,X_1) \cup [X_1,X_2) \cup \ldots \cup [X_{N-1},X_N]$$

where

$$X_i = i/N + 2bi(i-N) \quad (0 \leq i \leq N).$$

If $t \in [X_{i-1},X_i)$, then R chooses the action $a = (X_{i-1}+X_i)/2$ in equilibrium.

Since S always wants a to be higher than R wants $(b > 0)$, one might expect that, if t is very close to 1, S will try to reveal the fact. We therefore investigate whether there is a self-signaling neologism $X = (y,1]$, where $y > X_{N-1}$.

For such an X to be self-signaling, it must be the case that when $t = y$, S is indifferent between using the neologism (thus inducing $a = [1+y]/2$) and using his equilibrium message (thus inducing $a = [1+X_{N-1}]/2$). This implies that y+b, S's ideal, is just half way between those two possible actions:

$$(1+y)/2 - [y+b] = y+b - (1+X_{N-1})/2$$

This implies that

$$1 + y - 4y - 4b + 1 + X_{N-1} = 0$$

$$\begin{aligned} 3y &= 2 - 4b + X_{N-1} \\ &= 2 - 4b + [N-1]/N - 2b(N-1) \\ &= 3- 1/N - 2b(N+1) \end{aligned}$$

To check whether this describes a value of y in the range $(X_{N-1},1)$, we write

$$y = 1 - [1/N+2b(N+1)]/3 < 1$$

and

$$\begin{aligned} 3(y-X_{N-1}) &= 3/N + 6b(N-1) - 1/N - 2b(N+1) \\ &= 2/N + 4bN - 8b \\ &= 2/N + 4b(N-2) \end{aligned}$$

Thus, if $N \geq 2$, $y \in (X_{N-1},1)$. If $N = 1$, then $y \in (X_{N-1}, 1)$ if and only if $b < 1/2$. Hence, if $b < 1/2$, there is <u>no</u> neologism-proof equilibrium in this example; if $b > 1/2$, the uncommunicative equilibrium is the unique neologism-proof equilibrium. (It is also the unique sequential equilibrium.)

Discussion of the Examples

Example 1: For the uncommunicative outcome to be a sequential equilibrium requires beliefs $p(m) \in [1/3,2/3]$ for all $m \in M^*$, whether or not used in equilibrium. We find such beliefs implausible because (i) it is implausible that all the messages in M^* will be used for meaningless babble; (ii) we assume that some message not used in equilibrium will have t_1 as focal meaning, i.e. will be something like "Really, t_1 has occurred"; and (iii) since such a message has a focal (literal) meaning that S would want believed by R if and only if it is true, we assume that R will believe it. Thus the possibility of "unauthorized" communication rules out the outcome in which R always takes action a_3: it is a sequential equilibrium but is not neologism-proof.

Example 2: The notion of neologism-proof equilibrium does not simply mean selecting a most-communicative sequential equilibrium, as one might suspect.

Example 3:

We do not propose as a solution concept that type-1's can induce action a_1 and type-2's can induce action a_3. This is plainly not an equilibrium. The fact that the sequential equilibrium (in which all get a_3) is blocked by the ability of type-1's to get a_1 does not imply that there is an equilibrium in which they do so.

This language suggests a core-like approach to the problem. Type t_1 can get a payoff of 2 if he can separate himself. The usual core concept would automatically allow him to do so, so that the types will be treated separately. However, in our context, if pretending to be t_1 and being so treated is attractive to t_2, relative to his alternative, then he cannot be "kept out" by t_1.

In Myerson's [10] framework, the uncommunicative equilibrium in this game is a core mechanism. According to the concept of core mechanism, type-1 cannot say "I am of type 1; I propose that you take action a_1; notice that I would not wish to propose this if I were in fact of type 2": Myerson's requirement for credibility of a neologism insists that R's proposed action be a best response to his beliefs even if his beliefs do not shift as a result of the neologism.

Another way of seeing the contrast is as follows: In Myerson's framework, the neologism consists of proposing a new allocation that is an equilibrium if R makes inferences (in the appropriate direction) from the proposal, and is also an equilibrium if he does not. In our framework, the proposer need not worry about what happens if R fails to make the appropriate inference, because we assume that the message is so credible that R _will_ make the inference.

What will happen in this game, given that there is no convincing equilibrium? Clearly, we cannot give a convincing equilibrium argument for any outcome. However, it is interesting to see what happens if agents adjust at finite speed towards best responses. See Section 5 below.

4. Existence Results

Examples 3 and 4 above show that we cannot generally expect a neologism-proof equilibrium to exist. In this section we provide some sufficient conditions for existence.

Our first result concerns games in which S's preferences over R's beliefs are unrelated to S's true type. This is often the case, for instance, in simple labor-market models in which private information concerns "ability." Formally, we say that _S's preferences are uniform_ if, for all t_1, $t_2 \in T$ and all $a_1, a_2 \in A$,

$$u^S(a_1, t_1) > u^S(a_2, t_1) \rightarrow u^S(a_1, t_2) > u^S(a_2, t_2) \qquad (4.1)$$

Since we assume that $a^*(X)$, R's best response to beliefs $\pi|X$, is unique for each non-empty subset $X \subseteq T$, (4.1) implies that T itself is the only possible self-signaling subset (since in equilibrium each type must be indifferent among all equilibrium messages). It follows that:

Proposition 2: If S's preferences are uniform, then the uncommunicative equilibrium is neologism-proof.

Our second result concerns games of pure conflict: that is, the case where S's and R's preferences over actions are always diametrically opposed: if, given t, S prefers a to b then R prefers b to a. (This includes the zero-sum case.) In this case, there can be no self-signaling subset relative to the no-communication equilibrium. To see this, let X be a self-signaling subset. Then we know that

$$v(X,t) > v(T,t) \quad \text{if } t \in X$$

By assumption, this implies:

$$u^R(a^*(X),t) < u^R(a^*(T),t) \quad \text{if } t \in X;$$

and hence R's expected payoff, given $t \in X$, is strictly lower from action $a^*(X)$ than from action $a^*(T)$. But this contradicts the definition of $a^*(X)$.

Hence we get:

Proposition 3: In a game of pure conflict, the uncommunicative equilibrium is neologism-proof.

(In [3], we also showed that every sequential equilibrium in a two-person zero-sum game has the same payoffs as the uncommunicative equilibrium.)

These two results concern cases in which there can be no important communication (although it is possible to have equilibria with "inessential" communication[6]). We now turn to the harder task of providing partial existence results when communication matters. First, we note that, in the case of a game in which S's and R's interests coincide completely, the fully-revealing equilibrium (which always exists) is necessarily neologism-proof. (We would like to be able to state that this is the unique neologism-proof equilibrium in such a game. While this is true when T has only one, two or three elements, it fails when T has four elements: see the Appendix.)

We now give some less intuitive conditions for existence. In the process, we see why equilibrium fails to exist in Example 3: this, and similar insight, is our motivation for inquiring into some intrinsically unappealing conditions.

We say that S's preferences are <u>quasi-linear</u> if, for all $t \in T$ and all non-empty $A, B \subseteq T$,

$$\min[v(A,t),v(B,b)] \leq v(A \cup B,t) \leq \max[v(A,t),v(B,t)] \qquad (4.2)$$

Notice that Example 3 violates this condition (look at $t = t_2$, $A = \{t_1\}$, $B = \{t_2\}$). The condition is a combination of quasi-concavity and quasi-convexity, hence the name.

<u>Proposition 4</u>: If S's preferences are quasi-linear, then either the uncommunicative sequential equilibrium is neologism-proof, or there is a communicative sequential equilibrium.

<u>Proof</u>: In uncommunicative equilibrium, t gets $v(T,t)$. Suppose X is a self-signaling neologism. Then by definition

$$\left.\begin{array}{ll} v(X,t) > v(T,t) & \text{if } t \in X \\ v(X,t) \leq v(T,t) & \text{otherwise.} \end{array}\right\} \qquad (4.3)$$

When $t \in X$, (4.3) and (4.2) imply that

$$v(T\backslash X,t) \leq v(T,t) \leq v(X,t) \qquad (4.4)$$

and when $t \in X' = T\backslash X$, they imply that

$$v(X',t) \geq v(T,t) \geq v(X,t) \qquad (4.5)$$

But (4.4) and (4.5) imply that there is a sequential equilibrium in which some equilibrium messages mean "$t \in X$" and all others mean "$t \in X'$." This proves Proposition 4.

We would like to be able to extend Proposition 4 by finding conditions under which, if a sequential equilibrium is not neologism-proof, then there exists a more-communicative sequential equilibrium. That would imply the existence of a neologism-proof equilibrium, given that T is finite. (Evidently, the conditions would have to be violated in Example 2.) However, I have been unable to find natural sufficient conditions for that extension. One result along these lines, but with very strong assumptions, is:

Proposition 5: Consider a sequential equilibrium in pure strategies which partitions T into $T_1 \cup \ldots \cup T_k$. Let X be a self-signaling subset wholly contained within one of the T_i: say i=1. Suppose that $X' = T_1 \backslash X$ is also self-signaling. Then there is a sequential equilibrium in pure strategies partitioning T into

$$T = X \cup X' \cup T_2 \cup \ldots \cup T_k.$$

Proof: We need to show three things:

(a) If $t \in X$, then $v(X,t) \geq v(X',t)$ and $v(X,t) \geq v(T_j,t)$, where $j>1$.

We have

$v(X,t) > v(T_1,t)$ since $t \in X$ and X is self-signaling;

$v(T_1,t) \geq v(X'\ t)$ since $t \notin X'$ and X' is self-signaling;

$v(T_1,t) \geq v(T_j,t)$ since $t \in T_1$, and we began at a sequential equilibrium.

(b) Similarly, we prove that if $t \in X'$, $v(X',t) \geq v(X,t)$ and $v(X',t) \geq v(T_j,t)$.

(c) If $t \in T_j$, then $v(T_j,t) \geq v(X,t)$ and $v(T_j,t) \geq v(X',t)$ follow immediately from the fact that X and X' are self-signaling.

This proves Proposition 5.

We can also prove another partial result to the effect that a self-signaling X contained in a partition set can be made part of a finer sequential equilibrium. For this, we assume quasi-linear preferences, and an extra condition that ensures that, if a vague lie is unprofitable, then so is a purportedly more precise lie. We defer the precise condition until the reader has seen what is needed.

Proposition 6: Consider a sequential equilibrium in pure strategies, so that T is partitioned into $T_1 \cup \ldots \cup T_k$, and, for each i, all types in T_i induce action $a^*(T_i)$. If there is a self-signaling neologism X which lies wholly within (say) T_1, and S's preferences are quasi-linear, then the following partition defines another sequential equilibrium in pure strategies:

$$T = X \cup (T_1 \backslash X) \cup T_2 \cup \ldots \cup T_k.$$

provided that an extra condition is satisfied (see below).

<u>Proof</u>: We need to check that none of three classes of types will want to lie: those in X, those in T_1 but not in X, and those in some T_j, $j \neq 1$.

(a) If $t \in X$, then $v(X,t) > v(T_1,t) \geq v(T_j,t)$ for $j \neq 1$. Hence t will not claim falsely to be in T_j, $j \neq 1$. Also, since preferences are quasi-linear and $v(X,t) > v(T_1,t)$, we have $v(X',t) \leq v(T_1,t) < v(X,t)$, so that t will not falsely claim to be in X'.

(b) If $t \in X'$, then since preferences are quasi-linear and $v(X,t) \leq v(T_1,t)$, it follows that $v(X',t) \geq v(T_1,t)$. But since the original partition was a sequential equilibrium, $v(T_1,t) \geq v(T_j,t)$ for $j \neq 1$. Hence t will not falsely claim to be in T_j. Also, $v(X',t) \geq v(T_1,t) \geq v(X,t)$; so t will not falsely claim to be in X.

(c) If $t \in T_j$, then we know that

$$v(T_j,t) \geq v(T_1,t)$$

and that

$$v(X,t) \leq v(T_j,t)$$

We need to show that

$$v(X', t) \leq v(T_j, t)$$

Unfortunately, our assumptions so far do not imply this: it is possible that $t \in T_j$ would like to lie and claim to be in X'. This would be a somewhat "surprising" case: although T prefers to tell the truth than to tell the lie T_1, he strictly prefers a more precise lie $X' \subseteq T_1$. If this can be ruled out, then the proof of Proposition 6 is complete. A sufficient assumption is the following:

If $t \in T_j$ and $A \subseteq B$, and $B \cap T_j = \emptyset$, then either $v(A, t) \leq v(T_j, t)$ or $v(B, t) > v(T_j, t)$.

Next, we prove a result that guarantees, under a weaker assumption than quasi-linearity, that the full-revelation sequential equilibrium (if it exists) will be neologism-proof.

<u>Proposition 7</u>: If S's preferences are quasi-convex, i.e. if for all $t \in T$, $A, B \subseteq T$,

$$u^S(a^*(A \cup B), t) \leq \max[u^S(a^*(A), t), u^S(a^*(B), t)]$$

then no self-signaling set X can be a union of partition sets T_i. In particular, if full revelation is a sequential equilibrium, it is also neologism-proof.

Proof: Let $X = T_1 \cup \ldots \cup T_j \subseteq T = T_1 \cup \ldots \cup T_k$. Then by quasi-convexity applied (j-1) times

$$u^S(a^*(X),t) \leq \max[u^S(a^*(T_1),t),\ldots,u^S(a^*(T_j),t)]$$

so that X cannot be self-signaling.

But now if a sequential equilibrium in pure strategies partitions T into singletons, then there can be no self-signaling sets X.

Note: Observe how quasi-convexity is violated in Example 2 above.

A further existence result has been obtained by Joel Sobel (personal communication). This is phrased in terms of Myerson's [10] concept of a neutral optimum for a class of games which includes our simple communication games. Recall that a neutral optimum is a sequential equilibrium (Myerson uses more general terminology) which maximizes a weighted sum $\Sigma\lambda(t)v^S(t)$ among sequential equilibria, where the weights $\lambda(t) \geq 0$ must satisfy additional conditions which generically determine them. Sometimes in a neutral optimum some of the weights $\lambda(t)$ will be zero. Sobel defines a strict neutral optimum to be a neutral optimum that has $\lambda(t) > 0$ for all $t \in T$. Myerson proves the existence of a neutral optimum, but it may not be strict. Sobel shows that, if there is a strict neutral optimum, then it is neologism-proof. However, there is no reason to believe that a strict neutral optimum will "usually" exist.

These partial existence results do not imply that neologism-proof equilibrium will "usually" exist. What is one to make of that fact?

There is a core-like quality to the notion of neologism-proof equilibrium, as Myerson recognised in naming his related concept "core mechanisms." Types, or groups of types, can block an allocation if they can do better by distinguishing themselves, and if their distinguishing claim is credible. While coalitions in the usual core concept can keep out undesirable members, this is not possible here (the undesirable members can mimic anything said), so that the set of neologism-proof equilibria is not just the core of some related game. However, the non-existence problems have the same flavor, and in the same way they are not to be dismissed as exceptional.

What can we predict for games without equilibrium? One possibility is to look at how a reasonable dynamic process behaves over time. In Section 5, we explore that approach.

5. Evolution of Language and Lying

In this section we re-examine Example 3 above under the assumption that agents do not instantly optimize against their opponents' strategies, but rather move in that direction. We assume a finite speed of introduction of neologisms, and a finite rate of imitation by type-2's when enough type-1's are using a neologism that using the old strategy means being recognised as a type-2 (which, recall, gives the lowest payoff).

We suppose that M^* consists of the positive integers $\{1,2,3,...\}$. Initially, all types of S send message 1 all the time, and R always responds with action a_3. This is the (unique) sequential equilibrium.

Now, every k periods, the next neologism is discovered by a few ($f << 1$) of the type 1's, who then use it effectively to distinguish themselves from the other S's. We assume for simplicity that R's learning is instantaneous (very fast compared to how fast S can change). Thus, type-1's using a newly discovered neologism induce action a_1. Those using an older message induce R's best response to the mix of types using that message.

In each period, both type-1's and type-2's change their message-sending behavior in response to differences in payoffs. Thus, if old messages are inducing action a_3, then type-2's continue to send old messages, but type-1's move (at a finite rate) into sending a new message, if one is available. After enough type-1's have moved away from the old messages, those messages begin to induce action a_2 (since most of the S's sending them are now type-2's), so now it pays for the type-2's to begin to imitate the type-1's. Thus the old messages get abandoned, and the new ones become "discredited" by type-2's imitating type-1's (in preference to being identified as type-2's). For some while, perhaps, there is no effective communication. Then the next neologism is discovered, and communication by type-1's begins again.

We wrote a simple computer program (available on request) to simulate this process. The proportions of type-1's and of type-2's that get treated with each of R's three actions are plotted on Figure 1, for different values of k (the number of periods between neologisms), r (the rate at which type-1's change to better moves), and s (the corresponding rate for type-2's). In Table 1, summary statistics are given for various parameter values. In Table 2, we show the results of regressing the average proportions of the two types treated with the three actions on the parameter values.

As can be seen, increasing the "speed" parameters f, 1/k, r and s does not lead to a predominance of action a_3. The fully-optimizing model of game theory corresponds to all parameters being very large: but without more knowledge of how they compare, we cannot properly say that the uncommunicative equilibrium correctly describes limiting behavior as the parameters grow large.

6. Conclusion

In a sequential equilibrium, no agent can profitably lie. We introduced a plausible condition expressing the further equilibrium requirement that no agent can profitably deviate by coining a credible neologism. We showed that this rules out some implausible sequential equilibria; and that sometimes it rules out all sequential equilibria. We then simulated adaptive behavior in a game with no neologism-proof equilibrium, and found that the predictions of sequential equilibrium were not fulfilled.

Appendix: The Case $u^R \equiv u^S$

In this case, full revelation of t is a neologism-proof equilibrium, also Pareto dominates every other sequential equilibrium (generically, up to equivalence). But it need not be the only neologism-proof equilibrium.

If $|T| = 1$ or 2, it is easy to see that the full-revelation equilibrium is the only neologism-proof equilibrium. If $|T| = 3$, this is still true, as we will show, but if $|T| = 4$ then it need not be true.

For $|T| = 3$, suppose some other sequential equilibrium was neologism-proof. Each message used in equilibrium must be used with positive probability by all three types (otherwise some two types would find themselves in the case $|T| = 2$). By scaling payoffs, we can suppose that all three types are equally likely and that the equilibrium is uncommunicative. Normalize $u^S(a^*(T), t) = 0 = u^R(a^*(T), t)$ for all t.

We can assume that $u^S(a^*(\{t\}), t) > 0$ for all t: otherwise, some t would never appear in $P(X)$ for any X, and we would be reduced to the case $|T| = 2$. Normalize so that $u^S(a^*(\{t\}), t) = 1$ for all t.

Choose $t_1 \varepsilon T$. Write $a = u^S(a^*(t_1), t_2)$, $b = u^S(a^*(t_1), t_3)$. Since $\{t_1\}$ is not self-signaling, we know $\max(a, b) > 0$: say $a > 0$.

By definition of $a^*(T)$, we know $1 + a + b \leq 0$.

Now write
$$c = u^S(a^*(\{t_1, t_2\}), t_1)$$
$$d = u^S(a^*(\{t_1, t_2\}), t_2)$$
$$e = u^S(a^*(\{t_1, t_2\}), t_3).$$

Now since $\{t_1, t_2\}$ is not self-signaling, we know that <u>either</u> $e > 0$ <u>or</u> $\min(c, d) \leq 0$. If $\min(c, d) \leq 0$ then $c + d \leq 1 < 1 + a$; but this contradicts the definition of $a^*(\{t_1, t_2\})$, since it says that $a^*(\{t_1\})$ does better, given $\{t_1, t_2\}$, than $a^*(\{t_1, t_2\})$. If $\min(c, d) > 0$ and $e > 0$, then $c + d + e > 0$, so that $a^*(\{t_1, t_2\})$ does better than $a^*(T)$ given T: a contradiction. This proves the result for $|T| = 3$. Now we show by example that the uncommunicative equilibrium <u>can</u> be neologism-proof when $|T| = 4$. (This also shows, incidentally, that neologism-proof equilibrium need not be unique, since full-revelation is also neologism-proof.)

There are four equally likely types, labelled 1, 2, 3, 4. There are fifteen actions, each of which is the optimal action $a^*(X)$ for just one non-empty subset X of T. Payoffs can be described as follows:

Action: a = a*({ }):	Type 1	2	3	4
1	1	.1	-10	-10
2	-10	1	.1	-10
3	-10	-10	1	.1
4	.1	-10	-10	1
12	.6	.6	.1	-10
23	-10	.6	.6	.1
34	.1	-10	.6	.6
41	.6	.1	-10	.6
13	.6	-10	.6	.1
24	.1	.6	-10	.6
123	-.1	.9	.9	-10
234	-10	-.1	.9	.9
341	.9	-10	-.1	.9
412	.9	.9	-10	-.1
1234	0	0	0	0

It is easy to check that the uncommunicative equilibrium is neologism-proof. For example, the neologism $\{t_2\}$ fails because $P(\{t_2\}) = \{t_2, t_3\}$. The neologism $\{t_2, t_3, t_4\}$ fails because $P(\{t_2, t_3, t_4\}) = \{t_3, t_4\}$.

(One could imagine that, in fact, a neologism $\{t_2\}$ would be believed (even though not self-signaling) because R would think of S's alternative as including not only the equilibrium but also some other possible neologisms. This line of argument seems persuasive in this case, but is difficult to follow in general.)

Figure 1: Two graphs of the proportion of type-1's inducing action a_1. In the first, we see discontinuities of two types. The relatively small discontinuity at T = 250 is due to the introduction of a neologism (on the part of 8% of the type-1's). The other discontinuities, including the large one, result from a critical change in the mixture of types using a message. The downward discontinuity results from a neologism becoming discredited by the influx of type-2's.

In the second graph, we applied a smoothing algorithm and used a much smaller value of k. We see the oscillatory behavior clearly.

TABLE 1

Variable	Intercept	Log(K/40)	Log(r/.02)	Log(s/.02)	R^2
XL	.15	-.05	-.06	-.01	.18
	(35.9)	(-4.7)	(-5.8)	(-1.1)	
XM	.41	.17	-.05	.29	.54
	(47.6)	(8.7)	(-2.6)	(14.6)	
XH	.44	-.13	.11	-.28	.50
	(50.1)	(-6.1)	(5.3)	(-13.6)	
YL	.48	-.14	.20	-.36	.70
	(63.1)	(-8.1)	(11.3)	(-20.1)	
YM	.48	.10	-.25	.38	.70
	(59.5)	(5.1)	(-13.2)	(20.0)	
YH	.04	.05	.05	-.02	.51
	(22.9)	(10.9)	(11.0)	(-5.2)	

Explanation: The variable named e.g., XH denotes the fraction of type-1's (X's) treated as being of high (H) probability of being type 1 (and thus getting action a_1). Thus YM is the fraction of type 2's who induce action a_3 (corresponding to middling probability of type 1).

Parameters result from regression using 256 data points corresponding to k = 10, 20, 40, 80; f = .01, .02, .04, 08, r = .01, .02, .04, .08; s = .01, .02, .04, .08. Numbers under parameters are t-statistics.

K	F	R	S	FXL	FXM	FXH	FYL	FYM	FYH
7	.01	.30	.15	.15	.25	.60	.60	.25	.15
7	.01	.30	.30	.20	.61	.19	.41	.55	.04
7	.01	.30	.60	.09	.40	.51	.57	.43	.00
7	.10	.30	.15	.13	.24	.63	.60	.25	.15
7	.10	.30	.30	.18	.60	.22	.41	.55	.04
7	.10	.30	.60	.08	.42	.50	.56	.44	.00
10	.08	.01	.01	.20	.38	.42	.50	.50	.00
10	.08	.01	.02	.22	.40	.39	.46	.54	.00
10	.08	.01	.04	.12	.50	.38	.27	.73	.00
10	.08	.02	.01	.21	.28	.51	.64	.35	.00
10	.08	.02	.02	.25	.31	.44	.60	.38	.02
10	.08	.02	.04	.17	.43	.40	.39	.60	.01
10	.08	.04	.01	.16	.22	.63	.74	.26	.00
10	.08	.04	.02	.19	.25	.56	.69	.30	.01
10	.08	.04	.04	.25	.32	.44	.61	.36	.03
15	.01	.30	.15	.07	.70	.23	.37	.58	.05
15	.01	.30	.30	.10	.81	.09	.20	.78	.02
15	.01	.30	.60	.08	.59	.33	.39	.59	.02
15	.10	.30	.15	.09	.71	.20	.37	.58	.05
15	.10	.30	.30	.09	.81	.10	.20	.78	.02
15	.10	.30	.60	.08	.58	.34	.39	.58	.02
40	.08	.01	.01	.26	.31	.43	.60	.37	.03
40	.08	.01	.02	.15	.48	.37	.33	.65	.02
40	.08	.01	.04	.15	.45	.39	.35	.63	.03
40	.08	.02	.01	.17	.28	.55	.64	.31	.05
40	.08	.02	.02	.26	.34	.41	.58	.38	.03
40	.08	.02	.04	.16	.46	.38	.34	.63	.03
40	.08	.04	.01	.10	.23	.67	.71	.24	.05
40	.08	.04	.02	.15	.30	.54	.61	.32	.08
40	.08	.04	.04	.21	.53	.26	.45	.51	.04
50	.08	.01	.01	.25	.36	.40	.57	.39	.04
80	.08	.01	.01	.25	.34	.40	.58	.39	.03
80	.08	.01	.02	.15	.48	.37	.32	.65	.03
80	.08	.01	.04	.15	.48	.37	.32	.66	.02
80	.08	.02	.01	.15	.31	.54	.60	.32	.08
80	.08	.02	.02	.21	.53	.26	.44	.52	.04
80	.08	.02	.04	.10	.62	.28	.22	.73	.05
80	.08	.04	.01	.08	.24	.67	.61	.25	.14
80	.08	.04	.02	.11	.57	.32	.45	.48	.08
80	.08	.04	.04	.13	.74	.13	.27	.71	.02

TABLE 2

FOOTNOTES

1. With more than two players, communication becomes a much more complicated affair. For example, two players could whisper to each other, keeping information secret from a third. Or they could merely pretend to whisper to each other, without actually passing information. We assume that there are two players in order to avoid these problems: thus, whatever is announced is publicly heard.

2. (2.4) implies that the supremum in the definition of $v^S(t)$ is attained, which is not otherwise guaranteed since M^* is infinite. If $r(\bullet)$ were such that the supremum could not be attained, then S would have no best response, and $r(\bullet)$ could not be part of a sequential equilibrium.

3. Since the strategy spaces are infinite, the theorems of Kreps and Wilson [8] do not apply directly.

4. Let ρ permute M^*. Then if (s,p,r) is a sequential equilibrium, so is $(\rho_o s, p_o \rho^{-1}, r_o \rho^{-1})$, in obvious notation.

5. This is why we take M* to be infinite. With A and T finite, it is possible to show that every sequential equilibrium is

equivalent to one in which only finitely many messages are used.

6. For example, let $u^R(a,t) = 0 = u^S(a,t)$ for all $(a,t) \epsilon A x T$. Then any pattern of revelation is a neologism-proof equilibrium.

References

[1] Banks, J., and J. Sobel, "Equilibrium Selection in Signaling Games," mimeo, San Diego, 1985.

[2] Crawford, V., and J. Sobel, "Strategic Information Transmission," *Econometrica* *50* (1982), pp. 1431-1451.

[3] Farrell, J., "Communication Equilibrium in Games," mimeo, GTE Labs, 1985.

[4] Green, J., and N. Stokey, "A Two-Person Game of Information Transmission," mimeo, Harvard University, 1980.

[5] Grossman, S., and M. Perry, "Perfect Sequential Equilibrium," mimeo, Chicago, 1985.

[6] Kohlberg, E., and J.-F. Mertens, "On the Strategic Stability of Equilibria," mimeo, Harvard, 1984.

[7] Kreps, D., "Signaling Games and Stable Equilibria," mimeo, Stanford, 1984.

[8] Kreps, D., and R. Wilson, "Sequential Equilibria," *Econometrica 50* (1982), pp. 863-894.

[9] Lewis, D., *Convention*, Cambridge: Harvard University Press, 1969.

[10] Myerson, R., "Mechanism Design by an Informed Principal," *Econometrica* *51* (1983), pp. 1767-1797.

Bar Code On
Back Page